CLAIMS

WHAT IS CLAIMED IS:

1. A method comprising the steps of:

a. determining residue conservation scores for a plurality of reference residues;

b. identifying a cluster of connected reference residues;

c. determining the average residue conservation score of the residues that comprise said cluster;

d. determining the average residue conservation score of those residue that do not comprise said cluster; and

e. if the average determined in step c) is greater than the average determined in step d), selecting said cluster as a datum for one class of training data for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapted for determining a continuous SVM score of a cluster of residues on the surface of query protein.

2. The method of claim 1 wherein said residue conservation score is selected from the group consisting of the: residue conservation z-score; neighbor averaged residue conservation z-score; residue conservation p-score; or neighbor averaged residue conservation p-score.

3. A method comprising the steps of:

a. determining residue conservation scores for a plurality of query residues;

b. identifying a cluster of connected query residues;

c. determining the average residue conservation score of the residues that comprise said cluster;

d. determining the average residue conservation score of those residue that do not comprise said cluster; and

e. if the average determined in step c) is greater than the average determined in step d), selecting said cluster as a testing datum for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapted for determining a continuous SVM score for a cluster residues on the surface of a protein.

4. The method of claim 3 wherein said residue conservation score is selected from the group consisting of the: residue conservation z-score; neighbor averaged residue conservation z-score; residue conservation p-score; or neighbor averaged residue conservation p-score.

5. A method comprising the steps of:

a. identifying a void on the surface of a reference protein;

b. determining the volume of said void;

c. comparing the volume of said void to the volume of a water molecule; and

d. if the volume of said void is greater than the volume of a water molecule, selecting said cluster as a datum for one class of training data for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapted for determining a continuous SVM score of a cluster of residues on the surface of a protein.

6. A method comprising the steps of:

a. identifying a void on the surface of a query protein;

b. determining the volume of said void;

c. comparing the volume of said void to the volume of a water molecule; and

d. if the volume of said void is greater than the volume of a water molecule, selecting said cluster as a testing datum for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapated for determining an a continuous SVM score of a cluster of residues on the surface of a protein.

7. A method for comprising the steps of:

a. determining a three dimensional Delaunay tessilation of all or substantially of the reference residues of a reference structure based upon their three-dimensional coordinates;

b. determining the Alpha Shape of the reference residues from the Delaunay tessellation; and

c. identifying empty, connected Delaunay tetrahedrons, thereby identifying a void;

d. determining the volume of said void by summing the volume of the empty, connected Delaunay tetrahedrons determined in step c); and

e. if the volume of said void is greater than the volume of a water molecule, selecting said cluster as a datum for one class of training data for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapted for determining a

continuous SVM score of a cluster of residues on the surface of a query protein.

8. A method comprising the steps of:

a. determining a three dimensional Delaunay tessilation of all or substantially of the query residues of a query structure based upon their three-dimensional coordinates;

b. determining the Alpha Shape of the query residues from the Delaunay tessellation; and

c. identifying empty, connected Delaunay tetrahedrons, thereby identifying a void;

d. determining the volume of said void by summing the volume of the empty, connected Delaunay tetrahedrons determined in step c); and

e. if the volume of said void is greater than the volume of a water molecule, selecting said cluster as a testing datum for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapted for determining a continuous SVM score of a cluster of residues on the surface of a query protein.

9. A method comprising the steps of:

a. determining residue conservation scores and surface orientation scores for a plurality of the residues on the surface of a reference protein;

b. identifying a cluster of connected reference residues;

c. determining the average residue conservation score of the residues that comprise said cluster;

d. determining the average residue conservation score of those residue that do not comprise said cluster; and

e. if the average determined in step c) is greater than the average determined in step d) and if the distribution of the surface orientation scores that characterize the residues that comprise the cluster indicates that the cluster is concave, selecting said cluster as a datum for one class of training data for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapted for determining a continuous SVM score of a cluster of residues on the surface of a query protein.

10. The method of claim 9 wherein said residue conservation score is selected from the group consisting of the: residue conservation z-score; neighbor averaged residue conservation z-score; residue conservation p-score; or neighbor averaged residue conservation p-score.

11. A method comprising the steps of:

a. determining residue conservation scores and surface orientation scores for a plurality of the residues on the surface of a query protein;

b. identifying a cluster of connected query residues;

c. determining the average residue conservation score of the residues that comprise said cluster;

d. determining the average residue conservation score of those residue that do not comprise said cluster; and

e. if the average determined in step c) is greater than the average determined in step d) and if the distribution of the surface orientation scores that characterize the residues that comprise the cluster indicates that the cluster is concave, selecting said cluster as a testing datum for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapted for determining a continuous SVM score of a cluster of residues on the surface of a query protein.

12. The method of claim 11 wherein said residue conservation score is selected from the group consisting of the: residue conservation z-score; neighbor averaged residue conservation z-score; residue conservation p-score; or neighbor averaged residue conservation p-score.

13. A method comprising the steps of:

a. determining residue conservation scores and surface orientation scores for a plurality of the solvent accessible residues on the surface of a reference protein;

b. determining the statistical distribution of the surface orientation scores;

c. determining the putative functional residue limit based upon the statistical distribution of surface orientation scores;

d. determining a first surface orientation score threshold and a first residue conservation score threshold;

e. identifying those residues that are characterized by residue conservation scores that are greater than the first residue conservation score threshold and surface orientation scores that are greater than the first surface

orientation score threshold as putative functional residues, thereby determining the first pass putative functional residues;

f. identifying at least one cluster comprising connected first pass putative functional residues;

g. for each cluster which was identified, determining whether the number of first pass putative functional residues in the cluster exceed the putative functional residue limit; if it does not, selecting said cluster as a datum for one class of training data for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapted for determining a continuous SVM score of a cluster of residues on the surface of a query protein; otherwise determining a second surface orientation score threshold and a second residue conservation threshold score threshold;

h. identifying those residues that are characterized by residue conservation scores that are greater than the second residue conservation score threshold and surface orientation scores that are greater than the second surface orientation score threshold as putative functional residues, thereby determining the second pass putative functional residues;

i. identifying at least one cluster comprising connected second pass putative functional residues;

j. for each cluster which was identified, determining whether the number of second pass putative functional residues in the cluster exceed the putative functional residue limit; if it does not, selecting said cluster as a datum for

one class of training data for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapted for determining a continuous SVM score of a cluster of residues on the surface of a query protein; otherwise determining a third surface orientation score threshold and a second residue conservation threshold score threshold; and

k. repeating steps h-j until no cluster may be identified that comprises more putative functional residues than the putative functional residue limit.

14. The method of claim 13 wherein said residue conservation score is selected from the group consisting of the: residue conservation z-score; neighbor averaged residue conservation z-score; residue conservation p-score; or neighbor averaged residue conservation p-score.

15. The method of claim 14 wherein said statistical distribution of surface orientation score is determined by a method comprising the steps of:

a. determining the range of statistical orientation scores determined in step b) of claim 13; and

b. partitioning the surface orientation scores among a plurality of statistical bins wherein the width of each statistic bin is a fraction of the range of the surface orientation scores determined in step a).

16. The method of claim 15 wherein said putative functional residue limit is determined from a method comprising the steps of:

a. selecting the statistical bin containing the greatest number of surface orientation scores among the statistical bins determined in step b) of claim 15 that are each centered about a concave surface orientation score; and

b. identifying the putative functional residue limit with the number of surface orientation scores contained within the statistical bin selected in step a).

17. A method comprising the steps of:

a. determining residue conservation scores and surface orientation scores for a plurality of the residues on the surface of a query protein;

b. determining the statistical distribution of the surface orientation scores;

c. determining the putative functional residue limit based upon the statistical distribution of surface orientation scores;

d. determining a first surface orientation score threshold and a first residue conservation score threshold;

e. identifying those residues that are characterized by residue conservation scores that are greater than the first residue conservation score threshold and surface orientation scores that are greater than the first surface orientation score threshold as putative functional residues, thereby determining the first pass putative functional residues;

f. identifying at least one cluster comprising connected first pass putative functional residues;

g. for each cluster which was identified, determining whether the number of first pass putative functional residues in the cluster exceeds the putative functional residue limit; if it does not, selecting said cluster as a testing

datum for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapted for determining a continuous SVM score of a cluster of residues on the surface of a query protein; otherwise determining a second surface orientation score threshold and a second residue conservation threshold score threshold;

h. identifying those residues that are characterized by residue conservation scores that are greater than the second residue conservation score threshold and surface orientation scores that are greater than the second surface orientation score threshold as putative functional residues, thereby determining the second pass putative functional residues;

i. identifying at least one cluster comprising connected second pass putative functional residues;

j. for each cluster which was identified, determining whether the number of second pass putative functional residues in the cluster exceeds the putative functional residue limit; if it does not, selecting said cluster as a testing datum for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapted for determining a continuous SVM score of a cluster of residues on the surface of a query protein; and

k. repeating steps h-j until no cluster may be identified that comprises more putative functional residues than the putative functional residue limit.

18. The method of claim 17 wherein said residue conservation score is selected from the group consisting of the: residue conservation z-score; neighbor averaged residue conservation z-score; residue conservation p-score; or neighbor averaged residue conservation p-score.

19. The method of claim 18 wherein said statistical distribution of surface orientation score is determined by a method comprising the steps of:

a. determining the range of statistical orientation scores determined in step b) of claim 16; and

b. partitioning the surface orientation scores among a plurality of statistical bins wherein the width of each statistic bin is a fraction of the range of the surface orientation scores determined in step a).

20. The method of claim 19 wherein said putative functional residue limit is determined from a method comprising the steps of:

a. selecting the statistical bin containing the greatest number of surface orientation scores among the statistical bins determined in step b) of claim 19 that are each centered about a concave surface orientation score; and

b. identifying the putative functional residue limit with the number of surface orientation scores contained within the statistical bin selected in step a).

21. A method comprising the steps of:

a. determining residue conservation scores for a plurality of residues on the surface of a reference protein;

b. identifying a void on the surface of a reference protein;

c. determining the average of the residue conservation scores for the residues that comprise the void identified in step b);

d. determining the average residue conservation scores for the remaining residues that do not comprise the void identified in step b);

e. determining the volume of said void; and

f. if the volume of said void is greater than the volume of a water molecule and the average residue conservation score determined in step c) is greater than the average residue conservation score determined in step d), selecting said cluster as a datum for one class of training data for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapted for determining a continuous SVM score of a cluster of residues on the surface of a query protein.

22. The method of claim 21 wherein said residue conservation score is selected from the group consisting of the: residue conservation z-score; neighbor averaged residue conservation z-score; residue conservation p-score; or neighbor averaged residue conservation p-score.

23. A method comprising the steps of:

a. determining residue conservation scores for a plurality of residues on the surface of a query protein;

b. identifying a void on the surface of a query protein;

c. determining the average of the residue conservation scores for the residues that comprise the void identified in step b);

d. determining the average residue conservation scores for the remaining residues that do not comprise the void identified in step b);

e. determining the volume of said void; and

f. if the volume of said void is greater than the volume of a water molecule and the average residue conservation score determined in step c) is greater than the average residue conservation score determined in step d), selecting said cluster as a testing datum for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapted for determining a continuous SVM score of a cluster of residues on the surface of a query protein.

24. The method of claim 23 wherein said residue conservation score is selected from the group consisting of the: residue conservation z-score; neighbor averaged residue conservation z-score; residue conservation p-score; or neighbor averaged residue conservation p-score.

25. A method comprising the steps of:

a. determining residue conservation scores for a plurality of residues on the surface of a reference protein;

b. determining a three dimensional Delaunay tessilation of all or substantially of the reference residues of said reference structure based upon their three-dimensional coordinates;

c. determining the Alpha Shape of the reference residues from the Delaunay tessellation; and

d. identifying empty, connected Delaunay tetrahedrons, thereby identifying a void;

e. determining the average of the residue conservation scores for the residues that comprise the void identified in step d);

f. determining the average residue conservation scores for the remaining residues that do not comprise the void identified in step d);

g. determining the volume of said void by summing the volume of the empty, connected Delaunay tetrahedrons determined in step d); and

h. if the volume of said void is greater than the volume of a water molecule and the average residue conservation score determined in step e) is greater than the average residue conservation score determined in step f), selecting said cluster as a datum for one class of training data for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapted for determining a continuous SVM score of a cluster of residues on the surface of a query protein.

26. The method of claim 25 wherein said residue conservation score is selected from the group consisting of the: residue conservation z-score; neighbor averaged residue conservation z-score; residue conservation p-score; or neighbor averaged residue conservation p-score.

27. A method comprising the steps of:

a. determining residue conservation scores for a plurality of residues on the surface of a query protein;

b. determining a three dimensional Delaunay tessilation of all or substantially of the query residues of said query structure based upon their three-dimensional coordinates;

c. determining the Alpha Shape of the query residues from the Delaunay tessellation; and

d. identifying empty, connected Delaunay tetrahedrons, thereby identifying a void;

e. determining the average of the residue conservation scores for the residues that comprise the void identified in step d);

f. determining the average residue conservation scores for the remaining residues that do not comprise the void identified in step d);

g. determining the volume of said void by summing the volume of the empty, connected Delaunay tetrahedrons determined in step d); and

h. if the volume of said void is greater than the volume of a water molecule and the average residue conservation score determined in step e) is greater than the average residue conservation score determined in step f), selecting said cluster as a testing datum for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapted for determining a continuous SVM score of a cluster of residues on the surface of a query protein.

28. The method of claim 27 wherein said residue conservation score is selected from the group consisting of the: residue conservation z-score; neighbor averaged residue

conservation z-score; residue conservation p-score; or neighbor averaged residue conservation p-score.

29. A method comprising the step of selecting a validated functional cluster as a testing datum for use in a binary classification model adapted for identifying a cluster of functional residues on the surface of a query protein or adapted for determining a continuous SVM score of a cluster of residues on the surface of a query protein.

30. The method of claim 29 wherein said validated functional cluster is determined by identifying those residues in a protein-ligand structure whose solvent accessible surface area increases upon removal of the ligand.

31. A method for identifying a putative functional cluster comprising the steps of:

a. determining residue conservation scores and surface orientation scores for a plurality of the residues on the surface of a query protein;

b. identifying a cluster of connected query residues;

c. determining the average residue conservation score of the residues that comprise said cluster;

d. determining the average residue conservation score of those residue that do not comprise said cluster; and

e. if the average determined in step c) is greater than the average determined in step d) and if the distribution of the surface orientation scores that characterize the residues that comprise the cluster indicates that the cluster is concave, identifying said cluster as a putative functional cluster.

32. The method of claim 31 wherein said residue conservation score is selected from the group consisting of the: residue conservation z-score; neighbor averaged residue

conservation z-score; residue conservation p-score; or neighbor averaged residue conservation p-score.

33. A method for identifying a putative functional cluster comprising the steps of:

a. determining residue conservation scores and surface orientation scores for a plurality of the residues on the surface of a query protein;

b. determining the statistical distribution of the surface orientation scores;

c. determining the putative functional residue limit based upon the statistical distribution of surface orientation scores;

d. determining a first surface orientation score threshold and a first residue conservation score threshold;

e. identifying those residues that are characterized by residue conservation scores that are greater than the first residue conservation score threshold and surface orientation scores that are greater than the first surface orientation score threshold as putative functional residues, thereby determining the first pass putative functional residues;

f. identifying at least one cluster comprising connected first pass putative functional residues;

g. for each cluster which was identified, determining whether the number of first pass putative functional residues in the cluster exceeds the putative functional residue limit; if it does not, identifying such a cluster as a putative functional cluster; otherwise determining a second surface orientation score threshold and a second residue conservation threshold score threshold;

h. identifying those residues that are characterized by residue conservation scores that are greater than the second residue conservation score threshold and surface orientation scores that are greater than the second surface orientation score threshold as putative functional residues, thereby determining the second pass putative functional residues;

i. identifying at least one cluster comprising connected second pass putative functional residues;

j. for each cluster which was identified, determining whether the number of second pass putative functional residues in the cluster exceeds the putative functional residue limit; if it does not, identifying such a cluster as a putative functional cluster; otherwise determining a third surface orientation score threshold and a second residue conservation threshold score threshold; and

k. repeating steps h-j until no cluster may be identified that comprises more putative functional residues than the putative functional residue limit.

34. The method of claim 33 wherein said residue conservation score is selected from the group consisting of the: residue conservation z-score; neighbor averaged residue conservation z-score; residue conservation p-score; or neighbor averaged residue conservation p-score.

35. The method of claim 34 wherein said statistical distribution of surface orientation score is determined by a method comprising the steps of:

a. determining the range of statistical orientation scores determined in step b) of claim 33; and

b. partitioning the surface orientation scores among a plurality of statistical bins wherein the width of each statistic bin is a fraction of the range of the surface orientation scores determined in step a).

36. The method of claim 35 wherein said putative functional residue limit is determined from a method comprising the steps of:

a. selecting the statistical bin containing the greatest number of surface orientation scores among the statistical bins determined in step b) of claim 33 that are each centered about a concave surface orientation score; and

b. identifying the putative functional residue limit with the number of surface orientation scores contained within the statistical bin selected in step a).

37. A method for determining a putative functional cluster comprising the steps of:

a. determining residue conservation scores for a plurality of residues on the surface of a query protein;

b. identifying a void on the surface of a query protein;

c. determining the average of the residue conservation scores for the residues that comprise the void identified in step b);

d. determining the average residue conservation scores for the remaining residues that do not comprise the void identified in step b);

e. determining the volume of said void; and

f. if the volume of said void is greater than the volume of a water molecule and the average residue conservation score determined in step c) is greater than the average residue conservation score determined in step d), identifying said void as a putative functional cluster.

38. The method of claim 37 wherein said residue conservation score is selected from the group consisting of the: residue conservation z-score; neighbor averaged residue conservation z-score; residue conservation p-score; or neighbor averaged residue conservation p-score.

39. A method for determining a putative functional cluster comprising the steps of:

a. determining residue conservation scores for a plurality of residues on the surface of a query protein;

b. determining a three dimensional Delaunay tessilation of all or substantially of the query residues of said reference structure based upon their three-dimensional coordinates;

c. determining the Alpha Shape of the query residues from the Delaunay tessellation; and

d. identifying empty, connected Delaunay tetrahedrons, thereby identifying a void;

e. determining the average of the residue conservation scores for the residues that comprise the void identified in step d);

f. determining the average residue conservation scores for the remaining residues that do not comprise the void identified in step d);

g. determining the volume of said void by summing the volume of the empty, connected Delaunay tetrahedrons determined in step d); and

h. if the volume of said void is greater than the volume of a water molecule and the average residue conservation score determined in step e) is greater

than the average residue conservation score determined in step f),

identifying said void as a putative functional cluster.

40. The method of claim 39 wherein said residue conservation score is selected from the group consisting of the: residue conservation z-score; neighbor averaged residue conservation z-score; residue conservation p-score; or neighbor averaged residue conservation p-score.

41. A method for identifying at least one cluster of functional residues on the surface of a query protein comprising the steps of:

a. identifying at least one validated functional cluster from at least one reference protein;

b. determining at least one putative functional reference cluster from at least one reference protein;

c. representing each validated functional cluster determined in step a) and each putative functional reference cluster determined in step b) with a functional annotation score of the same form;

d. identifying at least one putative functional cluster on the surface of a query protein;

e. representing each putative functional cluster determined in step d) with a functional annotation score of the same form as the functional annotation scores used to represent the putative functional reference clusters and the validated functional clusters in step c);

f. for each putative functional cluster identified in step d) comparing its functional annotation score determined in step e) to the functional annotation scores determined in step c); and

g. for each putative functional cluster identified in step d) determining whether it may be classified as a validated functional cluster, thereby identifying said putative functional cluster as a true functional cluster or whether it may be identified as putative functional reference cluster, thereby identifying said putative functional cluster as non-functional cluster, based upon the comparison made in step f).

42. The method of claim 41 wherein said functional annotation score is a one dimensional functional annotation score selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score.

43. The method of claim 42 wherein said functional annotation score is a multi-dimensional functional annotation score comprising one subscore selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster

maximum neighbor averaged residue conservation z-score, cluster averaged neighbor averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score.

44. The method of claim 42 wherein in said functional annotation score is a multi-dimensional functional annotation score comprising:

a. a first subscore that reflects the residue conservation of a putative functional reference cluster, putative functional cluster or a validated functional cluster; and

b. a second subscore that reflects a topographic aspect that may be associated with a concave putative functional reference cluster, putative functional cluster or a validated functional cluster.

45. The methods of claim 44 wherein said first subscore is selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor

averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score;

and said second subscore is selected from the group consisting of the: cluster volume, cluster surface area, cluster "mouth" area, cluster "mouth" circumference, and cluster depth.

46. The method of claim 43 wherein said functional annotation score is a four dimensional functional annotation score consisting of the: cluster maximum residue conservation z-score, cluster surface area, cluster "mouth" area, and cluster depth.

47. The method of claim 45 wherein said comparison performed in step f) is made using a method selected from the group consisting of: support vector machines, Bayesian methods, neural network methods, and decision tree methods.

48. The method of claim 43 wherein said comparison performed in step f) is made using a method selected from the group consisting of: support vector machines, Bayesian methods, neural network methods, and decision tree methods.

49. The method of claim 45 wherein said comparison performed in step f) is made using a method selected from the group consisting of: support vector machines, Bayesian methods, neural network methods, and decision tree methods.

50. A method for identifying at least one cluster of functional residues on the surface of a query protein comprising the steps of:

a. identifying at least one validated functional cluster from at least one reference protein;

b. identifying at least one putative functional reference cluster on the surface of at least one reference protein;

c. representing each validated functional cluster identified in step a) and each putative functional reference cluster identified in step b) with a functional annotation score of the same form;

d. identifying at least one putative functional cluster on the surface of a query protein using;

e. representing each putative functional cluster determined in step d) with a functional annotation score of the same form as the functional annotation scores used to represent putative functional reference clusters and validated functional clusters in step b);

f. using a support vector machine to determine a hyperplane that defines a first set of functional annotation scores that characterize the validated functional clusters determined in step a) and that defines a second set of functional annotation scores that characterize the putative functional reference clusters determined in step b) based upon the functional annotation scores determined in step c);

g. determining for each functional annotation score determined in step e) whether it falls into the first set of functional annotation scores determined in step f) or falls into the second set of functional annotation scores determined in step f); and

h. for each functional annotation score identified in step g) as falling into the into the first set of functional annotation scores corresponding to the validated

functional clusters, identifying the corresponding putative functional cluster as a functional cluster; for each functional annotation score identified in step g) as falling into the second set of functional annotation scores corresponding to the putative functional reference clusters, identifying the corresponding putative functional cluster as a non-functional cluster.

51. The method of claim 50 wherein said functional annotation score is a one dimensional functional annotation score selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score.

52. The method of claim 50 wherein said functional annotation score is a multi-dimensional functional annotation score comprising one subscore selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster

maximum neighbor averaged residue conservation p-score, cluster averaged neighbor averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score.

53. The method of claim 50 wherein in said functional annotation score is a multi-dimensional functional annotation score comprising:

a. a first subscore that reflects the residue conservation of a putative functional reference cluster, putative functional cluster or a validated functional cluster; and

b. a second subscore that reflects a topographic aspect that may be associated with a concave putative functional reference cluster, putative functional cluster or a validated functional cluster.

54. The methods of claim 53 wherein said first subscore is selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score;

and said second subscore is selected from the group consisting of: the: cluster volume, cluster surface area, cluster "mouth" area, cluster "mouth" circumference, and cluster depth.

55. The method of claim 50 wherein said functional annotation score is a four dimensional functional annotation score consisting of the: cluster maximum residue conservation z-score, cluster surface area, cluster "mouth" area, and cluster depth.

56. The method of claim 52 wherein said putative functional reference clusters are determined using the method of claim 16 and said putative functional clusters are determined using the method of claim 36.

57. The method of claim 55 wherein said putative functional reference clusters are determined using the method of claim 16 and said putative functional clusters are determined using the method of claim 36.

58. The method of claim 52 wherein said putative functional reference clusters are determined using the method of claim 25 and said putative functional clusters are determined using the method of claim 39.

59. The method of claim 55 wherein said putative functional reference clusters are determined using the method of claim 25 and said putative functional clusters are determined using the method of claim 39.

60. A method for determining a continuous SVM score for a putative functional cluster comprising the steps of:

a. identifying at least one validated functional cluster from at least one reference protein;

b. identifying at least one putative functional reference cluster on surface of a least one reference protein using the same method that was used to identify said putative functional cluster;

c. representing each validated functional cluster identified in step a) and each putative functional reference cluster identified in step b) with a functional annotation score of the same form, thereby forming two sets of functional annotation scores;

d. representing said putative functional cluster with a functional annotation score of the same form as the functional annotation scores used to represent the putative functional reference clusters and validated functional clusters in step c);

e. using a support vector machine to determine a hyperplane that divides the two set of functional annotation scores determined in step c); and

f. determining a function that monotonically scales with the distance between the said functional annotation score determined in step d) and the hyperplane determined in step e), thereby determining a continuous SVM score of said putative functional cluster.

61. The method of claim 60 wherein said functional annotation score is a one dimensional functional annotation score selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster

median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score.

62. The method of claim 60 wherein said functional annotation score is a multi-dimensional functional annotation score comprising one subscore selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score.

63. The method of claim 60 wherein in said functional annotation score is a multi-dimensional functional annotation score comprising:

a. a first subscore that reflects the residue conservation of a putative functional reference cluster, putative functional cluster or a validated functional cluster; and

b. a second subscore that reflects a topographic aspect that may be associated with a concave putative functional reference cluster, putative functional cluster or a validated functional cluster.

64. The methods of claim 63 wherein said first subscore is selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score;

and said second subscore is selected from the group consisting of: the: cluster volume, cluster surface area, cluster "mouth" area, cluster "mouth" circumference, and cluster depth.

65. The method of claim 60 wherein said functional annotation score is a four dimensional functional annotation score consisting of the: cluster maximum residue conservation z-score, cluster surface area, cluster "mouth" area, and cluster depth.

66. The method of claim 62 wherein said putative functional cluster is determined using the method of claim 36 and said putative functional reference clusters are determined using the method of claim 16.

67. The method of claim 65 wherein said putative functional cluster is determined using the method of claim 36 and said putative functional reference clusters are determined using the method of claim 16.

68. The method of claim 62 wherein said putative functional cluster is determined using the method of claim 39 and said putative functional reference clusters are determined using the method of claim 25.

69. The method of claim 65 wherein said putative functional cluster is determined using the method of claim 39 and said putative functional reference clusters are determined using the method of claim 25.

70. A method for determining a continuous SVM score for a putative functional cluster comprising the steps of:

a. identifying at least one validated functional cluster from at least one reference protein;

b. identifying at least one putative functional reference cluster on surface of a least one reference protein using the same method that was used to identify said putative functional cluster;

c. representing each putative validated functional cluster identified in step a) and each putative functional reference cluster identified in step b) with a functional annotation score of the same form, thereby forming two sets of functional annotation scores;

d. representing said putative functional cluster with a functional annotation score of the same form as the functional annotation scores used to represent the putative functional reference clusters and validated functional clusters in step c);

e. using a support vector machine to determine a hyperplane that divides the two set of functional annotation scores determined in step c); and

f. determining the distance between the said functional annotation score determined in step d) and the hyperplane determined in step e), thereby determining a continuous SVM score of said putative functional cluster.

71. The method of claim 70 wherein said functional annotation score is a one dimensional functional annotation score selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score.

72. The method of claim 70 wherein said functional annotation score is a multi-dimensional functional annotation score comprising one subscore selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor

averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score.

73. The method of claim 70 wherein in said functional annotation score is a multi-dimensional functional annotation score comprising:

a. a first subscore that reflects the residue conservation of a putative functional reference cluster, putative functional cluster or a validated functional cluster; and

b. a second subscore that reflects a topographic aspect that may be associated with a concave putative functional reference cluster, putative functional cluster or a validated functional cluster.

74. The methods of claim 73 wherein said first subscore is selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score;

and said second subscore is selected from the group consisting of: the: cluster volume, cluster surface area, cluster "mouth" area, cluster "mouth" circumference, and cluster depth.

75. The method of claim 70 wherein said functional annotation score is a four dimensional functional annotation score consisting of the: cluster maximum residue conservation z-score, cluster surface area, cluster "mouth" area, and cluster depth.

76. The method of claim 72 wherein said putative functional cluster is determined using the method of claim 36 and said putative functional reference clusters are determined using the method of claim 16.

77. The method of claim 75 wherein said putative functional cluster is determined using the method of claim 36 and said putative functional reference clusters are determined using the method of claim 16.

78. The method of claim 72 wherein said putative functional cluster is determined using the method of claim 39 and said putative functional reference clusters are determined using the method of claim 25.

79. The method of claim 75 wherein said putative functional cluster is determined using the method of claim 39 and said putative functional reference clusters are determined using the method of claim 25.

80. A method for determining a continuous SVM score for a putative functional cluster determined using the methods of claim 36 comprising the steps of:

a. identifying at least one validated functional cluster from at least one reference protein;

b. identifying at least one putative functional reference cluster on surface of a least one reference protein using the method of claim 16;

c. representing each putative validated functional cluster identified in step a) and each putative functional reference cluster identified in step b) with a functional annotation score of the same form, thereby forming two sets of functional annotation scores;

d. representing said putative functional cluster with a functional annotation score of the same form as the functional annotation scores used to represent the putative functional reference clusters and validated functional clusters in step c);

e. using a support vector machine to determine a hyperplane that divides the two set of functional annotation scores determined in step c); and

f. determining a continuous SVM score of said putative functional cluster according to Equation 8.

81. The method of claim 80 wherein said functional annotation score is a one dimensional functional annotation score selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score.

82. The method of claim 80 wherein said functional annotation score is a multi-dimensional functional annotation score comprising one subscore selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score.

83. The method of claim 80 wherein in said functional annotation score is a multi-dimensional functional annotation score comprising:

a. a first subscore that reflects the residue conservation of a putative functional reference cluster, putative functional cluster or a validated functional cluster; and

b. a second subscore that reflects a topographic aspect that may be associated with a concave putative functional reference cluster, putative functional cluster or a validated functional cluster.

84. The methods of claim 83 wherein said first subscore is selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor

averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score;

and said second subscore is selected from the group consisting of: the: cluster volume, cluster surface area, cluster "mouth" area, cluster "mouth" circumference, and cluster depth.

85. The method of claim 80 wherein said functional annotation score is a four dimensional functional annotation score consisting of the: cluster maximum residue conservation z-score, cluster surface area, cluster "mouth" area, and cluster depth.

86. A method for determining the probability that a putative functional cluster is functional comprising the steps of:

a. selecting a plurality of reference proteins, each comprising a validated functional cluster;

b. for each reference protein, identifying one or more reference functional clusters using the same method that was used to identify said putative functional cluster;

c. for each reference functional cluster that was identified in step b), determining a functional annotation score that characterizes it;

d. selecting a lower threshold score of at least 35% and not greater than 100% and an upper threshold score that is not greater than 65%;

e. determining the fraction of reference functional clusters identified in step b) that correctly correspond to validated functional clusters selected in step a) based upon the upper and lower threshold scores selected in step d) at each functional annotation score, for a plurality of functional annotation scores;

f. determining a functional annotation score of the same type as used in step c) that characterizes said putative functional cluster; and

g. identifying the probability that said putative functional cluster is functional with the fraction of reference functional clusters, each characterized by a functional annotation score that is equal to the functional annotation score of said putative functional cluster, that are correctly identified as corresponding to validated functional clusters in step e).

87. The method of claim 86 wherein said functional annotation score is a one dimensional functional annotation score selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score.

88. The method of claim 86 wherein said functional annotation score is a multi-dimensional functional annotation score comprising one subscore selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score.

89. The method of claim 86 wherein in said functional annotation score is a multi-dimensional functional annotation score comprising:

a. a first subscore that reflects the residue conservation of a putative functional reference cluster, putative functional cluster or a validated functional cluster; and

b. a second subscore that reflects a topographic aspect that may be associated with a concave putative functional reference cluster, putative functional cluster or a validated functional cluster.

90. The methods of claim 89 wherein said first subscore is selected from the group consisting of the: cluster maximum residue conservation z-score, cluster averaged residue conservation z-score, cluster median residue conservation z-score, cluster maximum neighbor averaged residue conservation z-score, cluster averaged neighbor

averaged residue conservation z-score, cluster median neighbor averaged residue conservation z-score, cluster maximum residue conservation p-score, cluster averaged residue conservation p-score, cluster median residue conservation p-score, cluster maximum neighbor averaged residue conservation p-score, cluster averaged neighbor averaged residue conservation p-score, and cluster median neighbor averaged residue conservation p-score;

and said second subscore is selected from the group consisting of: the: cluster volume, cluster surface area, cluster "mouth" area, cluster "mouth" circumference, and cluster depth.

91. The method of claim 86 wherein said functional annotation score is a four dimensional functional annotation score consisting of the: cluster maximum residue conservation z-score, cluster surface area, cluster "mouth" area, and cluster depth.

92. The method of claim 86 wherein said functional annotation score is a continuous SVM score determined using the method of claim 64.

93. The method of claim 86 wherein said functional annotation score is a continuous SVM score determined using the method of claim 74.

94. The method of claim 84 wherein said functional annotation score is a continuous SVM score determined using the method of claim 84.

95. A method for determining the probability that a putative functional cluster determined using the method of claim 36 is functional comprising the steps of:

a. selecting a plurality of reference proteins, each comprising a validated functional cluster;

b. for each reference protein, identifying one or more reference functional clusters using the method of claim 16;

c. for each reference functional cluster that was identified in step b), determining a functional annotation score that characterizes it;

d. selecting a lower threshold score of at least 35% and not greater than 100% and an upper threshold score that is not greater than 65%;

e. determining the fraction of reference functional clusters identified in step b) that correctly correspond to validated functional clusters selected in step a) based upon the upper and lower threshold scores selected in step d) at each functional annotation score, for a plurality of functional annotation scores;

f. determining a functional annotation score of the same type as used in step c) that characterizes said putative functional cluster; and

g. identifying the probability that the putative functional cluster is functional with the fraction of reference functional clusters, each characterized by a functional annotation score that is equal to the functional annotation score of said putative functional cluster, that are correctly identified as corresponding to validated functional clusters in step e).

96. The method of claim 95 wherein said functional annotation score is a continuous SVM score determined using the method of claim 66.

97. The method of claim 95 wherein said functional annotation score is a continuous SVM score determined using the method of claim 67.

98. The method of claim 95 wherein said functional annotation score is a continuous SVM score determined using the method of claim 76.

99. The method of claim 95 wherein said functional annotation score is a continuous SVM score determined using the method of claim 77.

100. The method of claim 95 wherein said functional annotation score is a continuous SVM score determined using the method of claim 82.

101. A computer system comprising:

a. a processor;

b. a memory;

c. programming for an operating system; and

d. programming for the method of claim 16.

102. A computer system comprising:

a. a processor;

b. a memory;

c. programming for an operating system; and

d. programming for the method of claim 36.

103. A computer system comprising:

a. a processor;

b. a memory;

c. programming for an operating system; and

d. programming for the method of claim 49.

104. A computer system comprising:

a. a processor;

b. a memory;

c. programming for an operating system; and

d. programming for the method of claim 57.

105. A computer system comprising:

a. a processor;

b. a memory;

c. programming for an operating system; and

d. programming for the method of claim 67.

106. A computer system comprising:

a. a processor;

b. a memory;

c. programming for an operating system; and

d. programming for the method of claim 77.

107. A computer system comprising:

a. a processor;

b. a memory;

c. programming for an operating system; and

d. programming for the method of claim 85.

108. A computer system comprising:

a. a processor;

b. a memory;

c. programming for an operating system; and

d. programming for the method of claim 91.

109. A computer system comprising:

a. a processor;

b. a memory;

c. programming for an operating system; and

d. programming for the method of claim 101.